Data Exchanges, XML, and why the exchange problem is still unsolved

Anthony W. Isenor
Defence R&D Canada – Atlantic
9 Grove Street, Dartmouth, Nova Scotia
Canada B2Y 3Z7

anthony.isenor@drdc-rddc.gc.ca

To understand the exchange of data between systems, we may first consider conceptual models for the exchange of data. The first model relies on a central data structure for passing data among nodes. This is the model commonly used in meteorology and oceanography communities. A second model is more formal, and relies on instances of a common data model. Nodes exchange data with an instance of a common database, with data replicated between the common instance databases. The third conceptual model deals with wrapper software that encapsulates the data asset. Applications query the data asset using an intermediate layer, sometimes called an integrator or mediator, to identify the required data asset. The mediator then deals with the critical data issues like consolidation of parameter codes, units, replicate data, metadata content and multiple structures. The resulting data is provided to the user as a coherent and internally consistent data set.

All of these models support data sharing between nodes. The ICES/IOC[1] Study Group on the Development of Marine Data Exchange Systems Using XML (SGXML) examined numerous issues that are important for the sharing of data [1]. In particular, SGXML examined issues related to metadata, parameter dictionaries and data placement in XML structures.

In terms of metadata, the SGXML reviewed numerous international metadata standards for use with oceanographic data. The SGXML contributed to the mapping between standards by developing mappings between the Marine Environmental Data Information (MEDI) referral catalogue system, ISO 19115 and the European Directory of Marine Environmental Data (EDMED). These mappings are important to allow systems the ability to convert metadata records from one standard to another. This will be very important when combining data assets, each using a different metadata standard, or when conversion is required for utilization.

The SGXML also investigated the issue of parameter dictionaries. SGXML contributed to the development of the BODC[2] Parameter Dictionary. This is evident by the BODC dictionary population increase from 7982 entries in May 2002 to 14431 entries in May 2004. SGXML is also responsible for an in depth mapping between BODC and IFREMER[3] dictionaries and BODC and the DONAR/WADI (The Netherlands) data models. Perhaps more importantly, these mappings have continued in other projects and now encompass about 11 dictionaries in total.

The SGXML also made a contribution in the area of XML data structures. One effort resulted in the development of the Keeley Bricks [2]. The initial concept for the generic structures was based on the work of J. Robert Keeley (Marine Environmental Data Service, MEDS) in the 1980s. The initial idea recognized that many data types being delivered to the data centre contained information parts that were consistent across the data types. It was thought that these

---

[1] ICES – International Council for the Exploration of the Sea
IOC – Intergovernmental Oceanographic Commission
[2] BODC - British Oceanographic Data Centre
[3] IFREMER - Institut Francais pour le Recherche et l'Exploitation de la Mer

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **27 SEP 2005**	2. REPORT TYPE	3. DATES COVERED **00-00-2005 to 00-00-2005**
4. TITLE AND SUBTITLE **Data Exchanges, XML, and why the exchange problem is still unsolved**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defence Research and Development Canada - Atlantic,PO Box 1012, Dartmouth,Nova Scotia B2Y 3Z7, Canada,**		8. PERFORMING ORGANIZATION REPORT NUMBER **; DRDC-Atlantic-SL-2005-21**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) **DRDC-Atlantic-SL-2005-21**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

consistent parts could be formalized into structures, or Bricks. The formal Bricks could then be arranged in multiple ways to address the many structures present in the various ocean data types.

This effort resulted in the identification of 20 Bricks. The Bricks cover aspects of oceanographic data types such as analysis methods, calibration, instrumentation, provenance, unit and variable definition. A single data structure was then developed from the bricks and was found to be capable of storing a diverse set of oceanographic data types including: profile data, current meter data, underway temperature-salinity data, water sample data, acoustic doppler current profiling data (both moored and shipboard) and biological net tow data.

A second data investigation utilized some of the ideas and methods discussed during the SGXML meetings, applying these ideas to the Tokyo Bay Environmental Information Center Project. An XML structure and supporting software was developed and used for data collection efforts that supported the monitoring of Tokyo Bay. This work also utilized components of the Geography Markup Language (GML).

Another GML related effort attempted to incorporate all of the Keeley Brick information into GML. This resulted in a somewhat complicated set of relationships between the Brick content and the GML structure. GML implementation requires an abstraction of oceanographic data types, and thus potentially introduces complications in terminology.

There are also efforts underway to integrate data systems within the oceanographic community. The JCOMM[4] Expert Team on Data Management Practices (ETDMP) is exploring issues related to the identification and aggregation of data sets [3]. A funded ETDMP project is developing a system based on the conceptual wrapper model. The system has multiple layers of data providers, integrators and user applications. Users define their requirements at the user application layer. The integrator layer then directs the queries to appropriate data providers. The data providers retrieve data from the local system, then sending the data back to the integrator layer. The integrator layer will deal with the issues of parameter codes, data replication, etc., and provide the user with a single data set from the multiple sources.

In terms of data semantics related to parameter usage vocabularies, the Marine Metadata Interoperability (MMI) project is making an important contribution to identifying the relationships between parameters in different dictionaries [4]. These dictionaries, which actually represent managed vocabularies, are being aligned and mapped into the Web Ontology Language (OWL) by the MMI project. The OWL implementation allows the searching and discovery of terms by examining up and down the hierarchy formed by the implementation. By doing so, the user has the ability to find previously unknown terminology in other dictionaries that match the search term. As well, tools being developed under MMI allow users to create and manage groups of terms for their particular needs. Thus, users may define groups of similar terms, from multiple dictionaries, that have particular meaning to the user.

In the data exchange process, there are many important issues. Some of the international efforts addressing particular exchange issues are described in this summary paper. In all of these efforts, the critical underlying issue is an understanding of the data content (Figure 1). The difficulty in understanding the content is often related to the supporting metadata. Often, the supporting metadata descriptions are incomplete or use varied semantic descriptions and different vocabularies. The assets are also highly distributed and stored in many different data structures

---

[4] JCOMM – Joint WMO/IOC Commission on Oceanography and Marine Meteorology
WMO – World Meteorological Organization

and software formats. All of these factors can contribute to the loss or misinterpretation of the data content. Only when data exchange is seamless from a semantic perspective, will the exchange problem truly be solved.

1. Isenor, Anthony W. and Roy K. Lowry. 2005. Final Report of the ICES/IOC Study Group on the Development of Marine Data Exchange Systems using XML. DRDC Atlantic ECR 2005-005. March 2005.

2. Isenor, Anthony W., J. Robert Keeley and Joe Linguanti. 2003. Developing an eXtensible Markup Language (XML) Application for DFO Marine Data Exchange via the Web. DRDC Atlantic ECR 2003-025. May 2003.

3. Mikhailov, Nikolay, Evgeny Vyazilov, Sergey Belov, and Sergey Sukhonosov. 2005. JCOMM/IODE ETDMP Pilot Project, The Technology Prototype for the End-To-End Marine Data Management. Basic Solution, Development Status and Use for Supporting the Marine Activity. Unpublished manuscript.

4. Marine Metadata Interoperability Project. 2005. See http://marinemetadata.org/

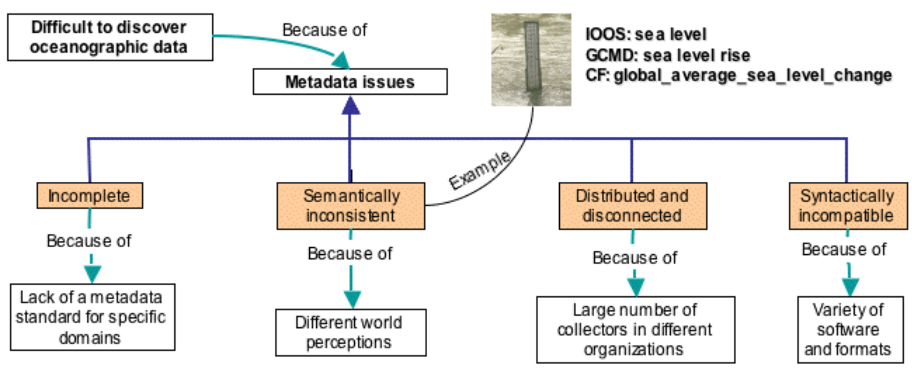

Figure 1: Schematic showing the difficulties associated the discovery process.

Image adapted from "HOW: Hydrologic Ontology for the Web". Luis Bermudez, Michael Piasecki, Dec, 2003. (AGU Poster.)